A Review of Comparisons of Aggregated Student Growth Percentiles and Value-Added for Educator Performance Measurement

Daniel F. McCaffrey and Katherine E. Castellano

Educational Testing Service

Acknowledgements: The brief benefitted greatly from helpful comments of J. R. Lockwood.

Abstract

Aggregate student growth percentiles and value-added offer two alternative methods to assess student achievement growth for teacher and school leader evaluations. States that are using student growth as part of teacher evaluations have split in their choice of which measure to use. Although both methods rely on comparing students' current achievement to their past achievement, the two approaches differ with respect to how they use the data to yield measures of educator performance. Georgia decided to use aggregated Student Growth Percentiles (AGP), either the mean (meanGP) or the median (medGP) of the Student Growth Percentiles of students linked to a teacher or leader. This brief reviews the literature on comparisons of AGP with value-added (VA) to assess the possible impact of Georgia's decision. The literature tends to find that AGP and VA are highly correlated and that a relatively small percentage of teachers change classification when AGP are substituted for VA measures. However, the teachers for which AGP and VA differ the most are not a random sample of all teachers. Several studies find that teachers of students with low prior achievement or from low-income families tend to rank relatively lower using AGP compared with other teachers than they would using VA, although two studies also find the opposite: VA had a stronger relationship to average prior achievement than did AGP. Only one study compared the inter-temporal stability of AGP and VA and found AGP to be more unstable across years, although the difference was not large. A second study found that use of the meanGP rather than the medGP would reduce statistical errors in AGP, which could improve year-to-year stability.

A Review of Comparisons of Aggregated Student Growth Percentiles and Value-Added for Educator Performance Measurement

The educator evaluation system in Georgia relies on multiple measures for teachers and leaders including measures of student growth. The system uses Student Growth Percentiles (SGPs) as the measure of student growth and aggregated SGPs (AGP) as the summary measure for teachers of students in grades and subjects tested by the standardized state tests. The SGP measures achievement growth by the percentile rank of a student's current achievement test score within the distribution of that student's peers who scored similarly on prior year tests. The mean or median SGP of students taught by a teacher or enrolled in a leader's school is used to summarize student growth and is the growth measure used in the Georgia educator evaluation systems. We use AGP as a general term for either the mean or median SGP, and when referring to the mean SGP specifically, we use "meanGP" and likewise use "medGP" to denote the median.

Alternative measures of student growth and educators' contributions to that growth also exist. These measures are commonly referred to as "value-added" (VA). A variety of statistical methods are used to calculate VA for teachers or leaders. All VA models and AGP share the common feature of using students' prior achievement to account for differences among students when measuring growth and comparing educators who teach students with different backgrounds. However, there are also differences between VA and AGP. For instance, VA models often also account for differences in demographics or economic status among classes or schools, and sometimes account for average classroom achievement or prior test scores from multiple subjects, whereas AGP, as currently used in practice, do not make any such adjustments.[1] In addition, VA uses linear models that rely on interval scaling of test scores but AGP avoid such assumptions.

Given the similarities and differences in the methods used to calculate VA and AGP, the Georgia Department of Education has asked how using VA instead of AGP might affect their educator performance measures. Several other authors have compared AGP and VA. This brief reviews those results and discusses their implications for Georgia educator measurement.

Background on AGP and VA Methodology

To facilitate the interpretation of the results of the comparison between VA and AGP, we start with a short review of commonly used methods in VA modeling and the calculation of SGP before turning to the literature review and the discussion of implications.

Aggregated Student Growth Percentiles Methods

AGP combine SGP to obtain a summary measure of student achievement growth for each teacher or leader. The SGP methodology uses a statistical model to calculate the percentiles of the distribution of current achievement scores for students with the same prior achievement history. Students with the same prior achievement history are considered peers for the calculation of SGP, so the percentiles characterize the distribution of current scores for a student's peers. The models for the percentiles do not assume that relationships between current and prior scores are linear. Each student's SGP is found by identifying the highest estimated percentile that a student's observed current score is greater than or equal to and rounding the corresponding percentile rank up to an integer value. For instance, a student whose observed

[1] The statistical models used to calculate SGP could incorporate additional student background variables. However, this is not the current practice used by any states implementing SGP and AGP and all of the studies reviewed in this brief consider only AGP calculated using the standard specification of including in the model only students' prior achievement in the same tested subject. Georgia is also using the standard specification to calculate SGP and AGP. We will consider only the standard specification to calculating SGP and AGP.

current score is greater than or equal to the estimated 51.5th percentile but less than the estimated 52.5th percentile receives an SGP of 52.

In current practice, only prior achievement scores from the same subject tests are used to match students with their peers via the SGP models. The statistical methods used to calculate SGP do not preclude the use of other variables, but the tradition has been to use only scores from tests in one subject area because the measures are meant to describe student achievement growth in a particular subject. SGP are calculated separately for each subject area. SGP implementations vary on the number of prior tests used in the calculation. Typical implementations use from one to three prior tests. Betebenner (2008) originally recommended using the median to estimate AGP due to the ordinal nature of percentile ranks, and medGP remain the most commonly used AGP, although meanGP are gaining wider use.

Value-Added Methods

Most of the approaches to VA modeling are variants of the same basic method. A statistical model is developed to predict a student's current achievement test scale score using that student's prior achievement test scores and other variables. Teachers or school leaders' VA scores are the average of the difference between their students' test scores and predicted scores from the model. The statistical model used to make the prediction is nearly always a linear regression model that assumes that the change in the predicted score for a unit change in a prior score is the same for all values of the prior score. The model relies on this same assumption for each of the variables used to make predictions.

One of the features that differentiate alternative VA methods is the terms included in the prediction model. All VA models include at least one prior test score. Some include multiple years of prior test scores on the same subject as the score being predicted, such as including the mathematics scores from the three most recent prior years when predicting a student’s current school year mathematics score. Other VA models include scores from other tested subjects; that is, including both prior year mathematics and reading scores when predicting reading or mathematics scores. VA models also often include additional student background variables in the prediction model. The most common variables are students’ free and reduced price lunch eligibility status (FRL) and demographics such as gender or race-ethnicity. Student English language proficiency and disability status are also very commonly used in the prediction models. Some models include variables such as students’ school attendance and disciplinary events. Different states and districts use various combinations of these variables for their models.

VA models might also account for the attributes of the students’ classroom or school peers or other characteristics of their classes or schools such as class size. Peer attributes are typically measured by averages of the student-level background variables, such as the classroom average prior score or the proportion of students with disabilities. All methods apply to schools and teachers; however, for clarity of presentation, in the remainder of this Value-Added Methods section, we describe VA for teachers.

A major distinction among VA methods is the use of a one-stage or two-stage modeling approach. Analysts carry out two-stage modeling much the same way they do the calculations of AGP. First, the analyst develops a prediction model for students’ current score using student background variables and prior achievement. No information about teachers is used in the prediction model—just like SGP. However, as noted previously, the prediction models used for VA typically include many variables not included in standard SGP calculations. In the second stage of two-stage VA, the analyst calculates the difference between students’ current and predicted scores and takes the average as the teachers’ VA. The mechanics of this VA and the

AGP approach are similar but differences remain between the methods, namely, the use of a linear model for this VA versus a nonlinear model for AGP, the possible inclusion of several student background variables for this VA versus the inclusion of only prior student tests for AGP, and the aggregation of student residuals (differences between observed and predicted student performance) for this VA versus the aggregation of student percentile ranks representing relative performance for AGP.

In a one-stage approach, the prediction model for generating VA treats the students' current scores as a function of the teachers and the students' background variables. The computations required to apply this model to the actual test score data return teacher VA along with the estimates of the other model parameters required to predict current student achievement scores. The teacher's VA still has the form of an average difference between the students' current scores and their predicted scores; but the computation is in a single process and so the common practice is to call this a one-stage approach. The most common method for implementing the one-stage approach is via teacher "fixed effects." The one-stage approach can also be implemented using teacher "random effects." The fixed-effects approach directly accounts for teachers when developing the prediction model, whereas the random effects approach accounts for teachers less directly. The two approaches often yield very similar VA measures, but will tend to differ the most for teachers with very few students used in their VA calculation. Moreover, the one-stage approaches (i.e., fixed and random effects) and the two-stage approach all make differential use of the variability of student background variables within and between classrooms when determining the prediction model, and this can lead to differences in the VA from the different approaches, including differences between the two one-stage approaches.[2]

Peer and other classroom or school-level predictors cannot be included in the prediction model in a one-stage approach calculated using teacher fixed effects and using only a single cohort of students for each teacher. To allow for inclusion of such variables, analysts typically use data from two or more cohorts of students from consecutive school years or use a two-stage approach.[3] When multiple cohorts of student data per teacher are used, each teacher's VA is assumed to be constant across all the cohorts included in the modeling effort.

A somewhat distinct VA approach is the EVAAS method that SAS uses in its calculations. Unlike the other VA methods, the EVAAS approach does not rely on linear regression to predict current achievement scores from prior scores and other background variables (Sanders, Saxton, and Horn, 1997). Rather it uses a multivariate model for all student scores in which student scores depend on their current and previous teachers and the scores from each student are correlated. The model, as implemented by SAS, does not include any adjustments for student background variables other than achievement scores.

[2] The fixed-effects approach models the relationships between background variables and test scores using only variation among students taught by the same teacher, "within teacher variance," and averages this variance across teachers. This is accomplished by including indicator variables for teachers in the VA model. The coefficients for these indicators serve as the teacher VA. The random-effects approach models the relationships between background variables and test scores using both the within teacher variance and the variation in scores and background variables between teachers, "between teacher variance." Data within teachers receives greater weight. As the number of students linked to teachers becomes larger, more weight is given to the pooled within teacher data and random-effects methods yield value-added measures very similar to those produced by fixed-effects methods. Random-effects methods also use "shrinkage" in the calculation of VA. Shrinkage tends to make the VA for teachers with few students less extreme and closer to the average measure for all teachers. The amount shrinkage affects teacher VA also decreases as the number of students linked to a teacher increase, which again contributes to random and fixed effects yielding similar VA, except for teachers with few students.

The two-stage approach ignores teachers during estimation of the prediction model and uses both within and between teacher variation to estimate model coefficients, but it places much greater weight on the between component than random-effects modeling. Because SGP calculations also ignore teachers, SGP also rely on the within and between variability in students prior achievement with weighting that is similar to the two-stage VA.

[3] Classroom-level variables can also be included in random-effects models, but this approach is not considered by any of the papers reviewed in this report.

Comparison of Value-Added and AGP

Seven studies have compared AGP and VA; four of them compared the measures for teachers and three compared the measures for schools. While the focus of this brief is on teachers, both measures can be used to evaluate schools as well. The research on schools is also informative for their use in teacher and leader evaluation.[4] Moreover, given the dearth of literature on the comparison of AGP and VA, it would seem amiss to ignore the studies that focused on schools instead of teachers. In the studies reviewed in this report, researchers compared different methods by computing the correlations between teacher (or school) AGP and VA from one or more models. Some of the studies also computed the correlation between teacher (or school) VA and AGP and their students' average demographic variables and prior achievement. It is desirable for this correlation to be small (close to zero). Some of the studies evaluated other technical properties as well, including the inter-temporal stability (i.e., how similar teacher [or school] measures are in consecutive years) for AGP and VA.

The comparisons of AGP and VA measures yield some common findings. First, AGP and VA scores for teachers are highly correlated. In the seven studies, the correlations between AGP and VA calculated for the same teachers ranged from .77 to .93 for teachers and .69 to .99 for schools. The researchers conducted the studies using data from different states, students at different grade-levels and a variety of VA modeling approaches, but they generally found similar results. The second common finding is that although the correlations are very high, AGP and VA differ notably for some teachers or schools. For instance, one study found that for 25% of

[4] Among the studies reviewed here the results for schools and teachers are often similar. The school studies typically consider only one subject area and sometimes focus on students of a particular grade-level within the school. In general, because schools have more students, school-level measures are typically more precise than teacher level measures and would tend to be more stable across years. However, the variability in teacher effectiveness tends to be greater than the variability in school effectiveness, and there is often less diversity of students within classrooms than within schools. These facts can offset the gains in precision due to larger sample sizes. Issues such as how the relationship between AGP and student background variables differs from the corresponding relationship for VA should be similar for schools and teachers.

teachers, the ranking by VA and AGP differed by 13 percentage points or more (Goldhaber, Walch, & Gabele, 2012). The third common finding is that the correlation between AGP and student background variables differed from the correlation between VA and background variables. Five studies discussed the relationship between aggregate student background data and AGP or VA. In three of those studies (Goldhaber, Walch, & Gabele, 2014; Ehlert, Kodel, Parsons, & Podgursky, 2012; Wright, 2010), AGP tended to show a stronger positive relationship with student prior achievement and a stronger negative relationship with the proportion of students from low income families than various VA measures. In one study, (Walsh & Isenberg, 2014), the opposite held: AGP had a weaker positive relationship with average prior achievement than VA did. Although, even in this study AGP had a stronger negative relationship with the percentage of English language learner (ELL) students in the classroom than did the study's VA methods. In the final study (Briggs & Betebener, 2009), the results were inconsistent. Differences between VA and AGP depended on the type of VA model, the included background variables, and the number of prior test scores included in that model. In the following subsections, that are ordered (reverse) chronologically starting with the most recent reports, we review and discuss the main findings of each study. Table A.1 in Appendix A provides a summary of the models compared in each study.

Walsh and Isenberg

Walsh and Isenberg (2014) compared AGP to VA measures calculated using a one-stage VA model. The study included mathematics and English Language Arts (ELA) scores for teachers of grade 4 to 8 students in Washington D.C. during the 2010-11 school year. They used medGP calculated with up to three prior years of scores. The VA measures were calculated using a fixed-effects, one-stage VA procedure that included students' immediately prior year mathematics and reading scores and background variables for race, FRL eligibility status, special education status, ELL status, and attendance in the prior school year.[5] The authors compared AGP to VA, investigated how using AGP rather than VA affects teacher classifications in the school district's IMPACT evaluation system, and studied factors related to the differences between AGP and VA.

IMPACT is the district's teacher performance measurement system. It provides teachers with a rating that classifies their performance into four categories on the basis of multiple performance measures which include student growth and classroom observations, similar to the Teacher Effectiveness Measure (TEM) score used by Georgia. Student growth accounts for 50% of the IMPACT rating.

Walsh and Isenberg found that AGP and VA measures correlated very highly: .93 for mathematics and .91 for reading. However, the authors pointed out that, even though the correlation is strong, AGP and VA measures would lead to different conclusions for some teachers. Specifically, the IMPACT classification would have changed for 14% of teachers in the district had it used AGP rather than VA in its teacher performance system.

In studying factors that were related to the differences between AGP and VA, the authors found that relative to VA, AGP were lower for teachers with higher proportions of ELL students. The authors explained this finding as a result of AGP not accounting for student language status. However, unlike most other studies, they also found that the relationship between AGP and student prior achievement scores was weaker than the corresponding relationship between VA and prior achievement. They conjectured that this second result could be an artifact of the SGP estimation method triggered by the matching of teachers to students so that teacher effectiveness is correlated with their students' prior achievement. For example, the proposed problem with

[5] The value-added model also controlled for measurement error in the prior year test scores.

SGP would arise if the more effective teachers were assigned to classrooms with higher average prior achievement scores and less effective teachers were assigned to classrooms with lower prior achievement.[6]

To explore the plausibility of this conjecture, they also calculated VA using a two-stage procedure. Recall, that like AGP, two-stage VA does not account for teachers when developing the prediction model and would be susceptible to similar errors due to this choice as AGP are. Walsh and Isenberg found that AGP do not favor teachers of classes with lower average prior achievement relative to this two-stage VA. They view this as support for their conjecture, but additional data and research are necessary to confirm it.

Guarino, Reckase, Stacy, and Wooldridge

Guarino et al. (2014) compared AGP to one-stage and two-stage VA using data for mathematics teachers of grades 5 and 6 students from a single large school district for the 2002 to 2007 school years.[7] The authors did not explicitly state which background variables were included in their VA specification, but they listed race or ethnicity, FRL and ELL status as variables in the data along with prior test scores. The authors also did not state explicitly how many years of prior test scores they used in their VA models. The text suggests that they did not include classroom aggregates of the student data in the two-stage model. The number of prior years of test scores used in the AGP calculations was also unspecified. The authors compared both meanGP and medGP to the VA measures. Although this report is lacking details, the results of their study reinforce the findings of the other studies in this review: AGP and VA yielded

[6] SGP adjust for student prior achievement using between classroom variation in student achievement. If teachers are matched to students so that teacher effectiveness is correlated (either positively or negatively) with student prior achievement, then using between classroom variation in student prior achievement when adjusting for prior achievement can conflate teacher effectiveness with the adjustment. This would result in more effective teachers receiving AGP that are relatively too low and less effective teachers receiving AGP that are relatively too high.

[7] The authors also compared AGP and VA using simulated data. Their simulated data analysis tests theoretical properties of the methods, but it does not describe their performance in practice so we do not discuss those results.

highly correlated measures. The correlations of one-stage VA with medGP or meanGP were .81 and .83 and the corresponding values for the two-stage VA were .77 and .79, respectively. The correlation of meanGP and medGP was .97. The finding that meanGP have stronger correlation with VA is consistent with the finding of Castellano and Ho (2014, discussed later) that medGP are less precise (have larger standard error) than meanGP. Lack of precision in a measure degrades its correlation with other measures (Allen and Yen, 1979).

Goldhaber, Walch, and Gabele

Goldhaber et al. (2014) used data from over 34,000 North Carolina teachers from 14 school years to compare various VA methods and medGP. The authors reported that medGP and meanGP were highly correlated so they only presented medGP results. They only included a single prior test score in their SGP models used to aggregate to medGP (and meanGP). They used a one-stage VA method with teacher fixed effects and used data from two adjacent school years (i.e., two cohorts of teacher data) to estimate value-added for each teacher for every two-year period. They considered two versions of this model. The first included individual student prior mathematics and reading test scores for the immediately prior year and background variables (gender, race or ethnicity, FRL status, learning disability status, ELL status, and parents' education levels). The second included all the variables used in the first VA specification as well as classroom percentages of FRL, disability, and minority students, percentage of students with parental education of bachelor's degree or higher, average prior year mathematics and reading achievement, and class size.[8, 9]

The authors found that VA and medGP correlated about .93 or .92 for mathematics

[8] The authors could include classroom level predictors in their one-stage model because they pooled together two years of data for teachers when calculating VA.

[9] Goldhaber et al. (2014) also considered a VA measure for teachers that relies only on within *school* variation among students. This method would not be used for teacher evaluations so we do not consider it in this review.

teachers and .83 or .84 for reading teachers, depending on the VA model used. The strong correlation did not guarantee the same rankings for most teachers. Twenty-five percent of teachers ranked in the bottom quintile of the distribution by VA (the lowest 20% of VA) received a higher ranking by AGP.[10] The same held for the top quintile (highest 20% of VA). Only about 50% of the teachers ranked in each of the second, third, or fourth quintiles by VA received the same ranking by AGP.

The authors also found that compared with VA, medGP have much stronger relationships with average classroom prior achievement, percentage of minority students and percentage of students eligible for subsidized school meals. For example, the authors identified classrooms as advantaged or disadvantaged. Advantaged classrooms are those with average student prior achievement in the highest quintile and with the proportion of FRL in the lowest quintile. Disadvantaged classrooms are those in the lowest quintile when ranked by average prior achievement and in the highest quintile when ranked by the proportion of FRL students. Using this classification, the authors find that average percentile ranks of the AGP for reading were 68 and 33, respectively, for teachers in advantaged classrooms and in disadvantaged classrooms, resulting in a difference of 35 percentile points. The comparable numbers for the VA model without classroom level variables were 58 for advantaged classrooms, 44 for disadvantaged classrooms, and a difference of 14 percentiles. For the VA model with classroom level predictors, these numbers were 60 for advantaged classrooms, 43 for disadvantaged classroom, for a 17 percentile point difference. Thus, the differences in teachers' average percentile ranks by AGP between advantaged and disadvantaged classrooms were higher than for either of the VA models. The pattern is the same for mathematics although the differences across methods were

[10] This result is for VA from the model without classroom-level variables. It is reported in another paper by Goldhaber on the same topic and using the same data (Goldhaber & Theobald, 2013).

smaller: The differences between advantaged and disadvantaged classroom were 9 and 14 percentile points for the two VA models and 21 for the AGP.

In a companion technical report (Goldhaber et al., 2012), the authors reported additional results of their analyses including results comparing AGP to VA from a model with only a single prior achievement score as a predictor. This VA model used exactly the same predictor variables as the SGP model used in calculating the AGP. The correlation between VA from this model and AGP was .94 for mathematics and .90 for reading. However, these VA measures were more strongly correlated with student background variables than VA from the other models: average percentile ranks for teacher in advantaged classrooms were 65 and 72 for mathematics and reading, and they were 38 and 29 for mathematics and reading for teachers in disadvantaged classrooms. The differences are 27 and 43 percentile points, respectively, which are more similar to those of the AGP than those of the other VA model.

In this study, the relationship between aggregate student variables and teacher measures (AGP or VA) seems to be more sensitive to the variables used in the modeling than whether or not a VA or AGP approach was used for calculating teacher measures. Consequently, the strong relationship between AGP and student background variables found by Goldhaber et al. (2014) may be inflated because the authors calculated AGP using only a single prior test score. The relationship may not have been as strong if the authors had used more years of prior test scores in their AGP; however, they did not report results for such models. Goldhaber et al. (2014) also compared the year-to-year stability in medGP to the stability of VA measures. They found that medGP are about 91% as stable across adjacent time periods as the VA measures for reading and about 94% as stable for mathematics.

Castellano and Ho

Castellano and Ho (2014) compared school-level meanGP and medGP to one stage VA calculated using school fixed effects, one-stage VA calculated using school random effects, and two-stage VA. They used three prior years of test scores for all methods and mathematics and reading test score data for grades 3 to 6 for about 550 schools. They found that the school-level meanGP and medGP were highly correlated with the three VA measures for both reading and mathematics. The correlations were over .93 for the medGP and over .98 for the meanGP. They also ranked schools using VA, from the fixed-effect model, and compared these rankings to rankings by either the meanGP or medGP for the reading data. They find that the median difference between VA rankings and meanGP rankings was 3 percentile points, and the median difference in rankings for VA and medGP was 6 percentile points. Moreover, the distributions of percentile differences between the VA and the medGP was much wider (maximum difference of about 40) than between the VA and the meanGP (maximum difference of about 25).

Unlike the other studies reviewed in this report, Castellano and Ho considered the standard error of AGP and directly assessed the size of the standard error in medGP relative to those of meanGP. They showed that the standard errors were about 70% larger for medGP than for meanGP. Using this number and theoretical results they determined that medGP requires about 3 times as many students in a group (teacher’s classroom or school) to obtain equal reliability as the meanGP.

A motivation for the use of AGP over VA is that the latter is more likely to suffer from bias if the assumption that test scores have interval scale properties is violated. Castellano and Ho (2014) directly assessed the sensitivity of the two methods to the test scale. They evaluated the extent to which meanGP and medGP and their one- and two-stage VA methods were invariant to transformations of the test score scale; that is, the extent the rank ordering of schools remained the same after applying various transformations to the test score scale. These transformations manipulated the distribution of each year's score—increasing the spread of score in some ranges and compressing the scores in other. The authors found that rankings by meanGP were the least sensitive to the scaling of the scores, rankings by medGP were second, followed by the rankings by the three VA measures, which were all very similar. These finding are consistent with expectations based on statistical theory.[11]

Ehlert, Koedel, Parsons, and Podgursky

Ehlert et al. (2012) used data from 1,846 schools enrolling students in grades 4 to 8 in Missouri to calculate two forms of VA measures and medGP for schools. The VA methods used linear regression to predict students' current year mathematics or ELA scores using both prior year mathematics and ELA scores and student level background data, including student race, gender, FRL status, ELL status, special education status, mobility status (mobile students are defined in the data as within-year building switchers) and grade-level. One of their VA methods was a one-stage approach, which used school fixed-effects, and the other was a two-stage approach that, in addition to the student-level variables used in the one-stage approach, also included school-level averages of these student-level variables.[12] The authors reported using as many prior year scores as were available for each student in calculating AGP, and they used scores from a single prior year for their VA. They pooled data from five years to create the school measures to reduce instabilities found in measures based on data from a single year.

Using these methods, the authors found that AGP correlated .82 with their one-stage VA and .85 with their two-stage VA. A correlation of .82 or .85 is high, but it leaves much room for

[11] See for example, Snedecor and Cochran (1980, p. 136).

[12] Ehlert et al. refer to their one-stage approach as "One-step VAM" for "one-step value-added model", and their two-stage approach as "Two-step VAM."

notable differences in the conclusions about school performance. The authors explored these differences and their relationship with the poverty level of the school's students. Specifically, they presented figures plotting the proportion of FRL students against the AGP or VA (for both of their VA methods). The figures show the strongest negative relationship for AGP, a similar but weaker negative relationship for the one-stage VA, and no relationship for the two-stage VA. They authors also studied how the different measures ranked high-poverty schools, where high-poverty schools are schools with at least 80% FRL students. They found that using AGP to rank schools results in high-poverty schools accounting for just 4% of schools ranked in the top quartile, whereas overall, 13% of schools in the state are high-poverty schools. Using their one-stage VA approach to rank schools resulted in high-poverty schools accounting for 10% of the top-quartile schools, and using their two-stage VA approach to rank schools resulted in high-poverty schools accounting for 15% of the top-quartile schools. Thus, even though AGP and VA generally ranked schools very similarly, schools that are ranked differently across the measures differ in their student populations, and AGP tends to give high-poverty schools lower ranks than either of the two alternative VA measures.

Wright

Wright (2010) compared medGP with several different VA measures. The analysis included one-stage VA models with teacher random effects and prior test scores but no teacher or school-level variables. Wright used four alternative specifications of one-stage VA with random teacher effects. The specifications differed only in the number of prior year tests included in the prediction model. The first specification used up to 12 tests (tests in four subjects from three prior years) when modeling students' current year mathematics scores, and the remaining models used 3, 2, or 1 prior year mathematics scores. All of the models included only prior achievement; no other student background variables were included. Wright also used the EVAAS model (Sanders et al., 1997) to calculate teacher VA. He compared these methods with three AGP specifications that used 3, 2, or 1 prior mathematics scores when calculating SGP.

Wright found that teacher scores from the EVAAS model had the weakest relationship with the proportion of the teacher's FRL students. The correlation between EVAAS VA scores and the proportion of FRL students equaled -.14 for grade 6 and -.03 for both grades 7 and 8. The correlation between AGP and the proportion of FRL students is similar to that of the one-stage VA measures with correlations ranging from about -.16 to -.38 depending on the students' grade level and the number of prior years of test scores used in the calculation. Correlations are stronger (further from zero) when fewer prior tests are used for either the VA or the AGP.

Briggs and Betebenner

Like Wright (2010), Briggs and Betebenner (2009) compared the medGP to the EVAAS method. They used reading scores for a single cohort of students followed for grades 3 (2003) to 6 (2006) who attended about 940 schools in the elementary grades (3 – 5) and about 640 schools in middle school (grade 6) in Colorado. They also considered only schools with at least 50 students in this cohort of interest, which reduced the number of schools by roughly two-thirds to about 570 grade 4 and 5 schools and 380 grade 6 schools. The results from the two samples were generally similar so we report only the results of the constrained sample. The medGP and EVAAS school effects were calculated for schools in grade 4 using 1 prior score, for grade 5 using 2 prior scores, and for grade 6 using 3 prior scores. Briggs and Betebenner directly compared medGP to EVAAS and compared their relationships with prior achievement and proportion of FRL students.

The authors found that medGP and EVAAS had moderate to strong relationships. Across grade levels, the correlations between medGP and EVAAS ranged from .72 to .91. They found correlations between medGP and mean prior achievement of .34 for grade 4 and .25 for grade 5. For EVAAS, the correlation with prior achievement was -.11 for grade 4 and -.09 for grade 5. However, for grade 6 schools, the correlation between EVAAS estimates of school effects and average prior achievement was .39, which is larger than the corresponding correlation of .31 for medGP. In grades 4 and 5, the correlations between medGP and the proportions of FRL students were -.42 and -.25, respectively. The corresponding values for EVAAS were about -.15 for grade 4 and -.05 for grade 5. This result is consistent with the results reported in Wright (2010): EVAAS had a weaker relationship with proportion FRL than did medGP. However, in grade 6, EVAAS had a stronger negative relationship with proportions of FRL students: the correlations were -.51 for EVAAS compared to -.39 for medGP. The authors provided no explanation for the differences between their results from grades 4 and 5 and those from grade 6.

Like Castellano and Ho (2014), Briggs and Betebenner also directly assessed the sensitivity of the two methods to the scale. They evaluated the extent to which medGP and EVAAS were invariant to transformations of the test that manipulated the shape of growth (i.e., linear versus nonlinear) and the variability of grade-to-grade score changes (i.e., increasing or constant standard deviations over grades). Consistent with the results of Castellano and Ho, Briggs and Betebenner also found that school rankings based on medGP were less sensitive to scale transformations than those based on a VA model, in this case, EVAAS.

The results of Briggs and Betebenner’s (2009) unpublished manuscript are included in this report for completeness. Unlike most of the other reports reviewed here, this paper has not undergone peer review by a journal. The lack of an investigation or explanation of the differences between results from grades 4 and 5 versus those from grade 6 is a weakness of the Briggs and Betebenner paper. Readers may want to consider these caveats as they evaluate the body of evidence from the collection of research studies presented in this literature review.

Implications for Georgia

Given the generally high levels of agreement between AGP and VA measures calculated using various specifications, the ranking of student growth measures for the majority of Georgia teachers would most likely not change substantially if the state used one of the common VA models rather than AGP. However, compared with AGP, VA might result in weaker correlations between the teacher (or school) measures and their students' average prior achievement, and the averages of other characteristics of the students in the schools or classrooms. Several of the studies found that AGP were more strongly correlated with average prior achievement or percentage of disadvantaged students than were VA. The level of change in the correlation will depend in part on the VA method used. Generally, in the reviewed literature, when VA models and AGP used similar variables in the modeling, they yielded measures with more similar correlations with student background variables. However, differences still existed between these methods even in these cases when they used the exact same variables, indicating that other differences in the methods also impact the teacher or leader growth measures. Moreover, common implementations of VA typically include multiple background variables not included in AGP and increasingly in research applications these include classroom aggregate variables through either a one-stage model using multiple years of data as in Goldhaber et al. (2014), or through a two-stage model as in Ehlert et al. (2012). Inclusion of these aggregate variables further reduces the correlation between VA and student background variables.

In the two studies that considered the EVAAS approach to VA, for the most part it yielded measures with weak correlations with background variables even though it includes only individual level student prior achievement in the models, like AGP. The results for EVAAS might be due to the assumptions which underlie the method.[13] If these assumptions are not consistent with the data, then this approach could spuriously suppress the correlation between VA and average prior achievement or the percentage of FRL students. This could be contributing to the low correlations found.

Another important consideration in an educator measure is how precisely it measures teacher's performance which is highly related to the year-to-year stability of the measures. Only Goldhaber et al. (2014) address this issue by comparing the inter-temporal stability of medSGP to VA, and they found that medGP were less stable across years. The work of Castellano and Ho (2014) clearly shows that meanGP are more precise than medGP, which would result in more year-to-year stability for meanGP. As stated above, based on statistical theory, this is expected.It is not clear how the stability of meanGP would compare to VA in the data used by Goldhaber et al.. This would be an area for future comparisons of the methods.

More generally, a limitation to this body of research is that nearly all the papers present results for only medGP and results might differ for meanGP since they will tend to be more precise, based on theory and as demonstrated by Castellano and Ho (2014). Guarino at al. (2014) did consider both meanGP and medGP and find very similar results except that meanGP tended to be more highly correlated with VA than medGP. This is to be expected since low precision suppresses correlation (Allen & Yen, 1979). Along the same lines, we could expect that meanGP would tend to have stronger correlation with background variables because of its greater

[13] See McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004), Lockwood, McCaffrey, Mariano, and Setodji (2007), and Mariano, McCaffrey, and Lockwood (2010) for discussion of the properties of the EVAAS model.

precision. The increase would be proportional to the increase found for the correlation with VA, so it would be small for the data used by Guarino et al. (2014).

These papers also only considered AGP when there was no correction for measurement error in the tests. However, Shang (2012), and Shang, Van Iwaarden, and Betebenner (2014) suggested using SIMEX to reduce potential bias in AGP due to measurement error in tests. Bias could create spurious correlation with student background variables so this approach might lead to different results than those presented in the papers reviewed here. A comparison of SIMEX-corrected AGP to VA could also be a useful topic for future comparisons of the methods.

The relative performance of AGP and VA is only one of many issues that would need to be considered in full comparison of the methods. A full understanding of the differences in the methods would require a review of the literature on the theoretical and empirical evaluations of the validity and reliability of both measures. Currently, such work has been done more fully for VA, so that method is somewhat better understood. Issues of implementation such as available software and data requirements would also need consideration. For example, software for calculating SGP is freely available and automatically accounts for missing prior test scores. Some software for fitting VA is proprietary and there are no packages designed to implement the one-stage or two-stage methods as conveniently as the SGP package. However, full consideration of these topics is beyond the scope of this literature review.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Betebenner, D. W. (2008). *A primer on student growth percentiles.* Retrieved from the Georgia Department of Education website: http://www.doe.k12.ga.us/

Briggs, D., & Betebenner, D. W. (2009, April). Is growth in student achievement scale dependent? In J. Doe (Chair) *Measuring and evaluating changes in student achievement: A conversation about technical and conceptual issues.* Invited symposium conducted at the meeting of the National Council on Measurement in Education, San Diego, CA. Retrieved from http://dirwww.colorado.edu/education/faculty/derekbriggs/Docs/Briggs_Weeks_Is%20Growth%20in%20Student%20Achievement%20Scale%20Dependent.pdf

Castellano, K. E., & Ho, A. D. (2014). *Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models* (Harvard University Technical Report). Retrieved from http://scholar.harvard.edu/files/andrewho/files/practical_differences_among_acsms_-_castellano_and_ho_2013.pdf (Note: This paper is forthcoming in the *Journal of Educational and Behavioral Statistics*. doi: 10.3102/1076998614548485)

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2012). *Selecting growth measures for school and teacher evaluations* (Working Paper 80). National Center for Analysis of Longitudinal Data in Educational Research. Retrieved from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtS

earch_SearchValue_0=ED535515&ERICExtSearch_SearchType_0=no&accno=ED535515 (Note: This paper is forthcoming in *Educational Policy*.)

Goldhaber, D, & Theobald. R. (2013, November) *Do different value-added models tell us the same things?* Carnegie Knowledge Network. Retrieved from http://carnegieknowledgenetwork.org/briefs/value-added/different-growth-models/

Goldhaber, D., Walch, J., & Gabele, B. (2012), *Does the model matter: Exploring the relationship between different student achievement-based teacher assessments* (CEDR Working Paper no. #2012-6). Retrieved from *http://www.cedr.us/papers/working/CEDR%20WP%202012-6_Does%20the%20Model%20Matter.pdf*

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy*, *1,* 28-39. doi:10.1080/2330443X.2013.856169

Guarino, C. M., Reckase, M. D., Stacy, B. W., & Wooldridge, J. M. (2014). *A comparison of growth percentile and value-added models of teacher performance* (Working Paper No. 39). Michigan State University, The Education Policy Center, East Lansing, MI. (Note: This paper is forthcoming in *Statistics and Public Policy*.)

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, *32*, 125-150. doi: 10.3102/1076998609346967

Mariano, L. T., McCaffrey, D. F., & Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, *35*, 253-279. doi: 10.3102/1076998609346967

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67-101. doi: 10.3102/10769986029001067

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Shang, Y. (2012). Measurement error adjustment using the SIMEX method: An application to student growth percentiles. *Journal of Educational Measurement*, *49*, 446–465. doi: 10.1111/j.1745-3984.2012.00186.x

Shang, Y., Van Iwaarden, A., & Betebenner, D. (2014, April). *Measurement error correction for the student growth percentile model*. Paper presented at the meeting of the National Council on Measurement in Education, Philadelphia, PA.

Snedecor, G. W., & Cochran, W. G. (1967). 1980. *Statistical methods*. Ames, IA: Iowa State University Press.

Walsh E., & Isenberg, E. (2014). *How does a value-added model compare to the Colorado growth model?* (Working Paper). Princeton, NJ: Mathematics Policy Research. Retrieved from http://mathematica-mpr.com/publications/pdfs/education/value_added_Colorado.pdf (Note: This paper is forthcoming in *Statistics and Public Policy*.)

Wright, S. P. (2010). *An investigation of two nonparametric regression models for value-added assessment in education* (White Paper). Cary, NC: SAS. Retrieved from http://www.sas.com/resources/whitepaper/wp_16975.pdf

Table A.1

Summary of Models Evaluated in the Studies Included in this Review

Authors (Year)	Method	Model Type	Variables	Notes
Walsh & Isenberg (2014)	Value-added	One-stage (fixed effects)	1 prior year math and reading scores, FRL, ELL, disability status, prior year attendance	The authors compared teachers
	AGP	medGP	3 prior year scores	
Goldhaber, Walch, & Gabele (2014)	Value-added	One-stage (fixed effects)	Prior year math and reading score, FRL, disability status, parental education, ELL, demographics	Authors used two years of data to calculate measures on teachers and considered additional models without providing full details in their paper
	Value-added	Two-stage	Same as one step plus classroom-level variables: class size, average prior year math and reading scores, %FRL, %parents with bachelors or higher, %disability, %minority	
	AGP	medGP	1 prior year score	
Castellano & Ho (2014)	Value-added	One-stage (fixed effects)	3 prior year scores	The authors compared schools using one data set and school districts with a second dataset; this review only considered the school results
	Value-added	One stage (random effects)	3 prior year scores	
	Value-added	Two-stage	3 prior year scores	
	AGP	meanGP	3 prior year scores	
	AGP	medGP	3 prior year scores	
Ehlert, Koedel, Parsons, & Podgursky (2012)	Value-added	One-stage (fixed effects)	One year prior math and ELA scores, demographics, FRL, ELL, mobility, disability status, grade-level, school averages of student level variables	Authors used five years of data to calculate measures on schools
	Value-added	Two-stage	Same as the one-stage model	
	AGP	medGP	Number of prior year scores used not reported	

Authors (Year)	Method	Model Type	Variables	Notes
Wright (2010)	Value-added	One-stage (random effects)	1, 2, or 3 prior math scores	Author compared teachers and considered additional methods not commonly used by education agencies and not discussed in this review
	Value-added	One-stage (random effects)	12 prior test scores from 4 subjects and 3 years	
	Value-added	EVAAS	12 prior test scores from 4 subjects and 3 years	
	AGP	medGP	1, 2, or 3 prior math scores	
Briggs & Betebenner (2009)	Value-added	EVAAS	1, 2, or 3 prior reading scores for grades 4, 5, and 6, respectively	The authors compared schools for one cohort and investigated scale-invariance of the methods that is not completely discussed in this review
	AGP	medGP	1, 2, or 3 prior reading scores for grades 4, 5, and 6, respectively	

Note. FRL is free or reduced price lunch (or meal) eligibility status, ELL is English language learner status.